The process of Machine Learning and using Overfitting to evaluate Linear Regression Model and Non-linear Regression

- Please compare the following two Regression Models to see which one has more serious overfitting issue.
  - Linear Regression Model 1
  - Non-Linear Regression Model 2
- Suppose we collect a set of sample data and distribute the sample data by
  - Training phase: 50%
  - Validation phase: 25%
  - Test phase: 25%

Training Phase			Validation Phase			Test Phase	
Real Data Set 1 50% of the collcted data	Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collcted data	Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collcted data	The better model (Model 1 or Model 2) selected from the Validation Phase based on the analysis of overfitting will be used to calculate ŷ

x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	x	ŷ=a1 + b1 * x or ŷ=a2 + b2 * x^2
1	1.8			1.5	1.7			1.4	
2	2.4			2.9	2.7			2.5	
3.3	2.3			3.7	2.5			3.6	
4.3	3.8			4.7	2.8			4.5	
5.3	5.3			5.1	5.5			5.4	
1.4	1.5			X	X	X	X	X	X
2.5	2.2			X	X	X	X	X	X
2.8	3.8			X	X	X	X	X	X
4.1	4.0			X	X	X	X	X	X
5.1	5.4			X	X	X	X	X	X

Note:

- Real Data Set 1 can be used to determine the formulas for Model 1: Linear Regression and Model 1: Linear Regression. That is, to determine the values of a1, b1, a2, and b2 in the following formulas:
  - y=a1 + b1 * x
  - y=a2 + b2 * x2
- After the formulas are determined, you can use the formulas to calculate the 欧 values in the following phases:
  - Training Phase
  - Validation Phase
  - Test Phase
- Optional: You may want to implement the following 3 programs:
  - Program 1: To implement Linear Regression Model 1
    Note:
    - This program is to use RealData Set 1 to determine a1 and b1 based on Model 1.
    - The program can be used to fill part of the blank spaces in above table.
  - Program 2: Non-Linear Regression Model 2

Note:

- This program is to use RealData Set 1 to determine a2 and b2 based on Model 2.
- The program can be used to fill part of the blank spaces in above table.

- Program 3: Calculate MSE

Answer:

Linear Regression

N=10,

x	y	x*y	x*x
1	1.8	1.8	1
2	2.4	4.8	4
3.3	2.3	7.59	10.89
4.3	3.8	16.34	18.49
5.3	5.3	28.09	28.09
1.4	1.5	2.1	1.96
2.5	2.2	5.5	6.25
2.8	3.8	10.64	7.84
4.1	4	16.4	16.81
5.1	5.4	27.54	26.01
∑x=31.8			
∑y=32.5			
∑xy=120.8			
∑x*x=121.34			

Slope(b) = (N∑XY - (∑X)(∑Y)) / (N∑X2 - (∑X)2)

B1= (10*120.8-31.8*32.5)/(10*121.34-31.8*31.8)=0.86

Intercept(a) = (∑Y - b(∑X)) / N

A1=(32.5-0.86*31.8)/10=0.52

Equation: y=0.52+0.86x

Non-Linear Regression

x	x^2	y	x*x*y	x^4
1	1	1.8	1.8	1
2	4	2.4	9,6	16
3.3	10.89	2.3	25.047	118.59
4.3	18.49	3.8	70.262	341.88
5.3	28..09	5.3	148.88	789.05
1.4	1.96	1.5	2.94	3.8416
2.5	6.25	2.2	13.75	39.063
2.8	7.84	3.8	29.792	61.466
4.1	16.81	4	67.24	282.58
5.1	26.01	5.4	140.45	676.62
∑x^2=121.34				
∑y=32.5				
∑x*x*y=509.76				
∑x^4=2330				

Slope(b) = (N∑XY - (∑X)(∑Y)) / (N∑X2 - (∑X)2)
B2= (10*509.76-121.34*32.5)/(10*2330-121.34^2)=0.13
Intercept(a) = (∑Y - b(∑X)) / N
A2=(32.5-0.13*121.34)/10=1.67
Equation: y=1.67+0.13x^2

Training Phase

Training Phase			
Real Data Set 1 50% of the collcted data		Model 1: Linear Regression	Model 2: Non-linear Regression
x	y	$y=a1 + b1 * x$	$y=a2 + b2 * x^2$
1	1.8	1.4	1.8
2	2.4	2.2	2.2
3.3	2.3	3.4	3.1
4.3	3.8	4.2	4.1
5.3	5.3	5.1	5.3
1.4	1.5	1.7	1.9
2.5	2.2	2.7	2.5
2.8	3.8	2.9	2.7
4.1	4	4	3.9
5.1	5.4	4.9	5.1
		y=0.52+0.86x	y=1.67+0.13x^2

Validation Phase

Validation Phase			
Real Data Set 2 25% of the collcted data		Model 1: Linear Regression	Model 2: Non-linear Regression
X	y	y=a1 + b1 * x	y=a2 + b2 * x2
1.5	1.7	1.8	2.0
2.9	2.7	3.0	2.7
3.7	2.5	3,7	3.4
4.7	2.8	4.6	4.5
5.1	5.5	4.9	5.1
		y=0.52+0.86x	y=1.67+0.13x^2

For MSE
Training Phase
Model 1
((1.4-1.8)^2+(2.2-2.4)^2+(3.4-2.3)^2+(4.2-3.8)^2+(5.1-5.3)^2+(1.7-1.5)^2+(2.7-2.2)^

2+(2.9-3.8)^2+(4-4)^2+(4.9-5.4)^2)/10=0.296

Model 2

((1.8-1.8)^2+(2.2-2.4)^2+(3.1-2.3)^2+(4.1-3.8)^2+(5.3-5.3)^2+(1.9-1.5)^2+(2.5-2.2)^2+(2.7-3.8)^2+(3.9-4)^2+(5.1-5.4)^2)/10=0.233

Validation Phase

Model 1

((1.7-1.8)^2+(2.7-3)^2+(2.5-3.7)^2+(2.8-4.6)^2+(5.5-4.9)^2)/5=1.028

Model 2

((1.7-2)^2+(2.7-2.7)^2+(2.5-3.4)^2+(2.8-4.5)^2+(5.5-5.1)^2)/5=0.792

Model 1: 1.028/0.296=3.472

Model 2: 0.792/0.233=3.399

Choose Model 2, because MSE is smaller

Test Phase

Use Model 2 for test phase

Test Phase	
Real Data Set 3 25% of the collcted data	The better model (Model 1 or Model 2) selected from the Validation Phase
x	y=a1 + b1 * x or y=a2 + b2 * x2
1.4	1.9
2.5	2.5
3.6	3.4
4.5	4.3
5.4	5.5
	y=1.67+0.13x^2

Final table

	Training Phase				Validation Phase				Test Phase	
	Real Data Set 1 50% of the collcted data		Model 1: Linear Regression	Model 2: Non-linear Regression	Real Data Set 2 25% of the collcted data		Model 1: Linear Regression	Model 2: Non-linear Regression	Real Data Set 3 25% of the collcted data	The better model (Model 1 or Model 2) selected from the Validation Phase
	x	y	y=a1 + b1 * x	y=a2 + b2 * x^2	x	y	y=a1 + b1 * x	y=a2 + b2 * x2	x	y=a1 + b1 * x or y=a2 + b2 * x2
	1	1.8	1.4	1.8	1.5	1.7	1.8	2.0	1.4	1.9
	2	2.4	2.2	2.2	2.9	2.7	3.0	2.7	2.5	2.5
	3.3	2.3	3.4	3.1	3.7	2.5	3.7	3.4	3.6	3.4
	4.3	3.8	4.2	4.1	4.7	2.8	4.6	4.5	4.5	4.3
	5.3	5.3	5.1	5.3	5.1	5.5	4.9	5.1	5.4	5.5
	1.4	1.5	1.7	1.9						
	2.5	2.2	2.7	2.5						
	2.8	3.8	2.9	2.7						
	4.1	4	4	3.9						
	5.1	5.4	4.9	5.1						
			y=0.52+0.86x	y=1.67+0.13x^2			y=0.52+0.86x	y=1.67+0.13x^2		y=1.67+0.13x^2

Github link:https://github.com/yinghe9999/Machine-Learning

Google Slides link:

https://docs.google.com/presentation/d/1riy-kBLNteD6goyh5ooQzrL0pIs4i5trzqptqk2C49Q/edit?usp=sharing